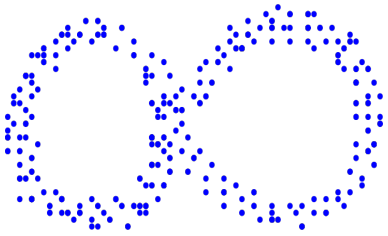
W dzisiejszych czasach dane wielkoskalowe stały się codziennością w wielu obszarach nauki i przemysłu, poczynając od internetu i mediów społecznościowych, przez genomikę, astrofizykę i nauki społeczne po projektowanie układów scalonych. Jak dotąd, analiza danych często jest wykonywana przez ludzi bazujących na swoim doświadczeniu, wspomaganym technikami i algorytmami statystycznymi. Niestety, przechowywane zbiory danych, często rozmiarów trudnych do wyobrażenia kilka dekad temu, są także często wysokowymiarowe, tzn. zawierają obiekty opisywane za pomocą setek, tysięcy lub nawet milionów parametrów. W takiej sytuacji dotychczasowe metody przestają wystarczać, co rodzi potrzebę rozwijania nowych algorytmów do analizy, wizualizacji i interpretacji danych.

W dziedzinie zwanej *wnioskowaniem geometrycznym (ang. geometric inference)* zbiór danych opisywany jest w języku geometrii, np. jako chmura punktów w przestrzeni geometrycznej. Celem jest pozyskanie informacji z danych korzystając z geometrii rozważanej przestrzeni oraz *wewnętrznej* geometrii danych. Dla przykładu, dla trójwymiarowej chmury punktów przedstawionej na rysunku 1 danej na wejściu, chcielibyśmy wywnioskować, że reprezentuje ona podwójny torus (∞). W powiązanej dziedzinie *topologicznej analizy danych* (TDA), na podstawie danych buduje się zagnieżdżone struktury topologiczne zwane *filtracjami kompleksów symplicjalnych* i oblicza ich homologie persystentne, co pozwala na wywnioskowanie informacji topologicznych. W obu tych dziedzinach znalazły zastosowanie zaawansowane techniki matematyczne, co doprowadziło w ciągu ostatnich dziesięciu lat do sukcesów w kilku obszarach zastosowań. Kolejną powiązaną dziedziną jest analiza dużych sieci, takich jak sieć WWW, czy sieci społecznościowe, o rozmiarach sięgających miliardów wierzchołków.



Rysunek 1: Chmura punktów w trzech wymiarach

Cel tego projektu jest dwojaki. Po pierwsze, chcemy połączyć koncepcje i narzędzia matematyczne z takich dziedzin jak rachunek prawdopodobieństwa, kombinatoryka, geometria dyskretna i różniczkowa a nawet geometryczna analiza funkcjonalna w celu uzyskania nowy efektywnych algorytmów, z dowodliwymi gwarancjami na czas działania i jakość generowanych wyników, dla problemów wnioskowania geometrycznego i topologicznej analizy danych. Wiele aktualnie używanych w TDA algorytmów działa w czasie zależnym wykładniczo od wymiaru danych, co nazywamy zjawiskiem *przekleństwa wymiaru*. W efekcie są one bezużyteczne już gdy liczba wymiarów sięga setek. Jednym z wyzwań jakie stawiamy przed sobą w tym projekcie jest pokazanie zastosowania randomizowanych technik redukcji liczby wymiarów, takich jak rzutowanie losowe, do problemów topologicznej analizy danych, takich jak obliczanie homologii persystentnych. Powiązanym kierunkiem badań jest rozszerzenie teorii wymiaru VC (Vapnika-Chervonenkisa) obejmującej pewne aspekty uczenia maszynowego i wnioskowania statystycznego, i zaadoptowanie jej do topologicznej analizy danych.

Drugim celem projektu jest odkrywanie geometrycznych i topologicznych własności danych, *w średnim przypadku*, poprzez badanie losowych modeli danych, z użyciem narzędzi probabilistycznych. Wspomniane modele losowe, takie jak grafy losowe, losowe kompleksy symplicjalne lub punkty na losowych wielościanach mogą dostarczyć wielu ciekawych intuicji na temat natury tych danych, a co za tym idzie efektywności metod obliczeniowych proponowanych dla TDA. Przykładowo, zamierzamy badać niedawno zaproponowaną procedurę *silnego kolapsu* zastosowaną na losowym kompleksie symplicjalnym. Kolejnym problem w tym obszarze jest badanie procesów perkolacyjnych, modelujących rozszerzanie się informacji, chorób lub plotek, w losowych lub deterministycznych kompleksach symplicjalnych.

Podsumowując, zamierzamy rozwinąć nowe techniki algorytmiczne oraz poprawić zrozumienie istniejących technik dla przetwarzania wysokowymiarowych zbiorów danych.